This application is submitted in the name of the following inventor:

Inventor	Citizenship	Residence (City and State)
Kleiman, Steven	United States	Los Altos, California

Title of the Invention

Scalable File Server with Highly Available Pairs

Background of the Invention

1. Field of the Invention

The invention relates to storage systems.

2. Related Art

Computer storage systems are used to record and retrieve data. One way storage systems are characterized is by the amount of storage capacity they have. The capacity for storage systems has increased greatly over time. One problem in the known art is the difficulty of planning ahead for desired increases in storage capacity. A related problem in the known art is the difficulty in providing scalable storage at a relatively ef-

ficient cost. This has subjected customers to a dilemma; one can either purchase a file system with a single large file server, or purchase a file system with a number of smaller file servers.

The single-server option has several drawbacks. (1) The customer must buy a larger file system than currently desired, so as to have room available for future expansion. (2) The entire file system can become unavailable if the file server fails for any reason. (3) The file system, although initially larger, is not easily scalable if the customer comes to desire a system that is larger than originally planned capacity.

The multi-server option also has several drawbacks. In systems in which the individual components of the multi-server device are tightly coordinated, (1) the same scalability problem occurs for the coordinating capacity for the individual components. That is, the customer must buy more coordinating capacity than currently desired, so as to have room available for future expansion. (2) The individual components are themselves often obsolete by the time the planned-for greater capacity is actually needed. (3) Tightly coordinated systems are often very expensive relative to the amount of scalability desired.

In systems in which the individual components of the multi-server device are only loosely coordinated, it is difficult to cause the individual components to behave in a coordinated manner so as to emulate a single file server. Although failure of a single

file server does not cause the entire file system to become unavailable, it does cause any files stored on that particular file server to become unavailable. If those files were critical to operation of the system, or some subsystem thereof, the applicable system or subsystem will be unavailable as a result. Administrative difficulties generally increase to due to a larger number of smaller file servers.

Accordingly, it would be advantageous to provide a method and system for performing a file server system that is scalable, that is, which can be increased in capacity without major system alterations, and which is relatively cost efficient with regard to that scalability. This advantage is achieved in an embodiment of the invention in which a plurality of file server nodes (each a pair of file servers) are interconnected. Each file server node has a pair of controllers for simultaneously controlling a set of storage elements such as disk drives. File server commands are routed among file server nodes to the file server node having control of applicable storage elements, and in which each pair of file servers is reliable due to redundancy.

It would also be advantageous to provide a storage system that is resistant to failures of individual system elements, and that can continue to operate after any single point of failure. This advantage is achieved in an embodiment of the invention like that described in co-pending Application Serial No. 09/037,652 filed March 10,1998, ~~Express Mail Mailing No. EE143637441US~~, in the name of the same inventor, titled

Available File Servers", ~~attorney docket number NAP-012,~~ US patent 6,317,944 hereby incorporated by reference as if fully set forth herein.

Summary of the Invention

The invention provides a file server system and a method for operating that system, which is easily scalable in number and type of individual components. A plurality of file server nodes (each a pair of file servers) are coupled using inter-node connectivity, such as an inter-node network, so that any one pair can be accessed from any other pair. Each file server node includes a pair of file servers, each of which has a memory and each of which conducts file server operations by simultaneously writing to its own memory and to that of its twin, the pair being used to simultaneously control a set of storage elements such as disk drives. File server commands or requests directed to particular mass storage elements are routed among file server nodes using an inter-node switch and processed by the file server nodes controlling those particular storage elements. Each file server node (that is, each pair of file servers) is reliable due to its own redundancy.

In a preferred embodiment, the mass storage elements are disposed and controlled to form a redundant array, such as a RAID storage system. The inter-node network and inter-node switch are redundant, and file server commands or requests arriving at the network of pairs are coupled using the network and the switch to the appropriate pair and processed at that pair. Thus, each pair can be reached from each other

pair, and no single point of failure prevents access to any individual storage element. The file servers are disposed and controlled to recognize failures of any single element in the file server system and to provide access to all mass storage elements despite any such failures.

Brief Description of the Drawings

Figure 1 shows a block diagram of a scalable and highly available file server system.

Figure 2A shows a block diagram of a first interconnect system for the file server system.

Figure 2B shows a block diagram of a second interconnect system for the file server system.

Figure 3 shows a process flow diagram of operation of the file server system.

Detailed Description of the Preferred Embodiment

In the following description, a preferred embodiment of the invention is described with regard to preferred process steps and data structures. However, those skilled in the art would recognize, after perusal of this application, that embodiments of the invention may be implemented using one or more general purpose processors (or special purpose processors adapted to the particular process steps and data structures) operating under program control, and that implementation of the preferred process steps and data structures described herein using such equipment would not require undue experimentation or further invention.

Inventions described herein can be used in conjunction with inventions described in the following applications:

- Application Serial No. 09/037,652, filed March 10, 1998, Express Mail Mailing No. EE143637441US, in the name of the same inventor, titled "Scalable and Highly Available File Server", ~~attorney docket number NAP-012.~~ US Patent 6,317,844

This application is hereby incorporated by reference as if fully set forth herein. It is herein referred to as the "Availability Disclosure."

File Server System

Figure 1 shows a block diagram of a scalable and highly available file server system.

A file server system 100 includes a set of file servers 110, each including a coupled pair of file server nodes 111 having co-coupled common sets of mass storage devices 112. Each node 111 is like the file server node further described in the Availability Disclosure. Each node 111 is coupled to a common interconnect 120. Each node 111 is also coupled to a first network switch 130 and a second network switch 130.

Each node 111 is coupled to the common interconnect 120, so as to be able to transmit information between any two file servers 110. The common interconnect 120 includes a set of communication links (not shown) which are redundant in the sense that even if any single communication link fails, each node 111 can still be contacted by each other node 111.

In a preferred embodiment, the common interconnect 120 includes a NUMA (non-uniform memory access) interconnect, such as the SCI interconnect operating at 1 gigabyte per second or the SCI-lite interconnect operating at 125 megabytes per second.

Each file server 110 is coupled to the first network switch 130, so as to receive and respond to file server requests transmitted therefrom. In a preferred embodiment there is also a second network switch 130, although the second network switch 130 is not required for operation of the file server system 100. Similar to the first network switch 130, each file server 110 is coupled to the second network switch 130, so as to receive and respond to file server requests transmitted therefrom.

File Server System Operation

In operation of the file server system 100, as further described herein, a sequence of file server requests arrives at the first network switch 130 or, if the second network switch 130 is present, at either the first network switch 130 or the second network switch 130. Either network switch 130 routes each file server request in its sequence to the particular file server 110 that is associated with the particular mass storage device needed for processing the file server request.

One of the two nodes 111 at the designated file server 110 services the file server request and makes a file server response. The file server response is routed by one of the network switches 130 back to a source of the request.

Interconnect System

Figure 2A shows a block diagram of a first interconnect system for the file server system.

In a first preferred embodiment, the interconnect 120 includes a plurality of nodes 111, each of which is part of a file server 110. The nodes 111 are each disposed on a communication ring 211. Messages are transmitted between adjacent nodes 111 on each ring 211.

In this first preferred embodiment, each ring 211 comprises an SCI (Scalable Coherent Interconnect) network according to IEEE standard 1596-1992, or an SCI-lite network according to IEEE standard 1394.1. Both IEEE standard 1596-1992 and IEEE standard 1394.1 support remote memory access and DMA; the combination of these features is often called NUMA (non-uniform memory access). SCI networks operate at a data transmission rate of about 1 gigabyte per second; SCI-lite networks operate at a data transmission rate of about 125 megabytes per second.

A communication switch 212 couples adjacent rings 211. The communication switch 212 receives and transmits messages on each ring 211, and operates to bridge messages from a first ring 211 to a second ring 211. The communication switch 212 bridges those messages that are transmitted on the first ring 211 and designated for

transmission to the second ring 211. A switch 212 can also be coupled directly to a file server node 110.

In this first preferred embodiment, each ring 211 has a single node 111, so as to prevent any single point of failure (such as failure of the ring 211 or its switch 212) from preventing communication to more than one node 111.

Figure 2B shows a block diagram of a second interconnect system for the file server system.

In a second preferred embodiment, the interconnect 120 includes a plurality of nodes 111, each of which is part of a file server 110. Each node 111 includes an associated network interface element 114. In a preferred embodiment, the network interface element 114 for each node 111 is like that described in the Availability Disclosure.

The network interface elements 114 are coupled using a plurality of communication links 221, each of which couples two network interface elements 114 and communicates messages therebetween.

The network interface elements 114 have sufficient communication links 221 to form a redundant communication network, so as to prevent any single point of

failure (such as failure of any one network interface element 114) from preventing communication to more than one node 111.

In this second preferred embodiment, the network interface elements 114 are disposed with the communication links 221 to form a logical torus, in which each network interface element 114 is disposed on two logically orthogonal communication rings using the communication links 221.

In this second preferred embodiment, each of the logically orthogonal communication rings comprises an SCI network or an SCI-lite network, similar to the SCI network or SCI-lite network described with reference to figure 2A.

Operation Process Flow

Figure 3 shows a process flow diagram of operation of the file server system.

A method 300 is performed by the components of the file server system 100, and includes a set of flow points and process steps as described herein.

At a flow point 310, a device coupled to the file server system 100 desires to make a file system request.

At a step 311, the device transmits a file system request to a selected network switch 130 coupled to the file server system 100.

At a step 312, the network switch 130 receives the file system request. The network switch 130 determines which mass storage device the request applies to, and determines which file server 110 is coupled to that mass storage device. The network switch 130 transmits the request to that file server 110 (that is, to both of its nodes 111 in parallel), using the interconnect 120.

At a step 313, the file server 110 receives the file system request. Each node 111 at the file server 110 queues the request for processing.

At a step 314, one of the two nodes 111 at the file server 110 processes the file system request and responds thereto. The other one of the two nodes 111 at the file server 110 discards the request without further processing.

At a flow point 320, the file system request has been successfully processed.

If any single point of failure occurs between the requesting device and the mass storage device to which the file system request applies, the file server system 100 is still able to process the request and respond to the requesting device.

- If either one of the network switches 130 fails, the other network switch 130 is able to receive the file system request and transmit it to the appropriate file server 110.

- If any link in the interconnect 120 fails, the remaining links in the interconnect 120 are able to transmit the message to the appropriate file server 110.

- If either node 111 at the file server 110 fails, the other node 111 is able to process the file system request using the appropriate mass storage device. Because nodes 111 at each file server 110 are coupled in pairs, each file server 110 is highly available. Because file servers 110 are coupled together for managing collections of mass storage devices, the entire system 100 is scalable by addition of file servers 110. Thus, each cluster of file servers 110 is scalable by addition of file servers 110.

- If any one of the mass storage devices (other than the actual target of the file system request) fails, there is no effect on the ability of the other mass storage devices

to respond to processing of the request, and there is no effect on either of the two nodes 111 which process requests for that mass storage device.

Alternative Embodiments

Although preferred embodiments are disclosed herein, many variations are possible which remain within the concept, scope, and spirit of the invention, and these variations would become clear to those skilled in the art after perusal of this application.